Keyword-Based Sentiment Mining using Twitter

Keyword-Based Sentiment Mining using Twitter

M. Baumgarten, Department of Computing & Engineering, University of Ulster, Newtownabbey, UK

M. D. Mulvenna, Department of Computing & Engineering, University of Ulster, Newtownabbey, UK

N. Rooney, Department of Computing & Engineering, University of Ulster, Newtownabbey, UK

J. Reid, RepKnight, Belfast, UK

ABSTRACT

Big Data are the new frontier for businesses and governments alike. Dealing with big data and extracting valuable and actionable knowledge from it poses one of the biggest challenges in computing and, simultaneously, provides one of the greatest opportunities for business, government and society alike. The content produced by the social media community and in particular the micro blogging community reflects one of the most opinion- and knowledge-rich, real-time accessible, expressive and diverse data sources, both in terms of content itself as well as context related knowledge such as user profiles including user relations. Harnessing the embedded knowledge and in particular the underlying opinion about certain topics and gaining a deeper understanding of the overall context will provide new opportunities in the inclusion of user opinions and preferences. This paper discusses a keyword-based classifier for short message based sentiment mining. It outlines a simple classification mechanism that has the potential to be extended to include additional sentiment dimensions. Eventually, this could provide a deeper understanding about user preferences, which in turn could actively and in almost real time influence further development activities or marketing campaigns.

Keywords: *Big Data, Data Sources, Keyword-Based Classifier, Micro Blogging Community, Sentiment Dimensions, Short Message Based Sentiment Mining, Social Media Community, User Relations*

INTRODUCTION

Today's connected society is characterized by the way people share information and by how such information affects the community as a whole. This is particular relevant when such information reflects the opinion of individuals about other individuals, companies, products, specific product features, etc. Arguably, Twitter is one of the most popular platforms for publishing opinions and other information to a global audience. In general, such platforms enable the networked community to easily express likes or dislikes, to convey personal feelings or moods, to comment about events or activities of other individuals, to publish news about general

DOI: 10.4018/jaci.2013040104

topics or themselves. Individually, they can be seen as simple statements that have limited or no value at all. However, collectively they reflect a powerful mechanism that can actively influence other people's behavior. For instance, if a large number of people approve of a given product by expressing their satisfaction with it then other people may be more likely to buy this particular product, which is referred to as online word-of-mouth branding. Vice versa, people may choose not to buy a certain product if a substantial amount of people has commented negatively about it. Equally compelling in this context is how individuals think about celebrities for which public opinion can be considered a valuable commodity. Disturbingly in both cases is that it is often irrelevant if the publicized opinions are based on actual facts or if they are based on unfounded information, hearsay, harassment, etc.

Moreover, individual comments are often simply aggregated into positive, negative and sometimes also into neutral categories that supposedly reflect the public's opinion. Such a simplification is not only problematic but it often leads to false interpretations because expressed sentiments seldom refers to a single general topic but generally to a highly specific context that is defined by multiple factors including time of occurrence, topicality, related topics, demographics, etc. Analyzing the sentiment individuals have towards a given context or person could not only help to assess the reputation of the context or person concerned but could also influence future decisions and actions. However, extracting the sentiment from relatively short and often slang-based messages such as tweets is a non-trivial task. Similar, correlating such sentiments to specific aspects, properties or even complex personal profiles is equally challenging

The following reviews the foundations and challenges of short message based sentiment analysis, briefly reviews the state of the art in this area and also prototypes a keyword based classifier which categorizes individual tweets into distinct groups that reflect different sentiments. Over time, this can be used to reason about the changing opinion a given context has within the connected community. Towards the end of this paper, a performance study is presented and future work is discussed before concluding remarks are given.

FOUNDATIONS AND CHALLENGES OF SENTIMENT ANALYSIS

Based on Wikipedia, sentiment analysis is defined as an "application of natural language processing, computational linguistics and text analytics" with the objective to identify and to "extract subjective information" from various content. In other words, sentiment analysis refers to the problem of extracting the opinion or emotion a person has about a given context, product, person, etc. While personal opinion can be expressed in various forms such as written, verbal as well as visual, it is, for the scope of this work, assumed that it is expressed or can be converted into short text fragments such as typically published via various micro blogging platforms or topic specific forums. Micro blogging platforms in general but Twitter in particular have become the platform of choice for the real-time publishing of personal messages that relate to products, celebrities, personal or public events or other contexts of interest. On an individual level they simply reflect the personal statement of a single person about a specific topic at a given moment in time. However, collectively they are expressive enough to reason about the social reputation a person or product may have at any moment in time and equally important in real-time. The most basic sentiment analysis is to determine the polarity a given message has in relation to a given topic. That is if the underlying sentiment is positive, negative or neutral. Other alignments such as angriness, sadness, happiness, etc. reflect more complex dimensions in this area of research that are more challenging to detect. One of the popular methods to determine the polarity is referred to as keyword spotting in which a given text is classified according to the occurrence

of relevant positive and negative keywords. Although a very simple yet effective method, it entails a number of drawbacks as summarized next. Firstly, keywords reflect unigrams, which makes it difficult to extract opinions or emotions from more complex structures. For instance, "A reader will find it difficult not to recommend this book" includes "recommend" which is normally regarded as a positive indicator with respect to the polarity of the phrase. However, because it is preceded by 'not', which indicates a negation of 'recommend', it should lead to a negative sentiment classification. Nevertheless, the first part of the above phrase "will find it difficult" negates this again pointing to a positive sentiment for the book concerned. This simple example clearly shows that a keyword based approach has some limitations to determine the polarity of complex phrases in particular if it includes sarcasm, irony, slang based expressions, misspellings, abbreviations etc. Nevertheless as shown later on, the accuracy of such approaches is still within acceptable limits. Other challenges in this context can be summarized as follows:

- **Polarity:** The polarity of a any specific piece of content is subject to interpretation and as such subjective. That is that it may differ for different contexts or persons and may even be topic and time specific. As such, any classification cannot unambiguously be regarded as a "gold standard". Instead it should be evaluated in relation to the context it is has been determined for;
- **Sentiment Variety:** Although, humans can express a multitude of sentiments most current algorithms only distinguish between positive and negative categories. Capturing additional dimensions is an important aspect in this area, as it would allow for a better and more detailed interpretation of the publics opinion;
- **Sentiment Strength:** Content is often categorized into distinct groups without quantifying the strength of the detected sentiment or the confidence of its correctness. However, in human terms the individual sentiments are expressed at various degrees ranging e.g. from very strong to very mild. Quantifying the strength a sentiment has for a given topic would allow for a better interpretation, which could also serve as a measure of confidence;
- **Topic Dependency:** A sentiment usually refers to a specific topic or feature. However, the topic may not necessarily be included in the same phrase as the sentiment which is particular relevant for micro blogs. Such mismatch can significantly influence the classification process significantly;
- **Multiple Topics:** Similar to the above, a given content may include multiple topics with potentially contradicting sentiments. The correlation of sentiments to the different and potentially related topics poses a significant challenge in this area;
- **Topic Dimensionality:** General topics are often composed or influenced by other sub-topics. Correlating and aggregating dependent content and the underlying sentiment together to evaluate the various aspects poses a significant challenging;
- **Noise:** Micro blogs are often created on mobile devices while on the move. As such micro blogs are usually not well formed or grammatically correct. Instead, they often include syntactic errors, are grammatically blurred and contain demographic or context specific abbreviations. Dealing with such noise is very difficult in general NLP but considering the small size of micro blogs this problem is even more significant;
- **Slang/Vernacular/Abbreviations:** Micro blogs often include user specific slang as well as abbreviations, which allow for shorter messages, personalized expressions, etc. Both of which are very much language dependent and may also differ demographically which makes it very difficult to correctly interpret them;
- **Sarcasm, Irony, Mockery, Cynicism, Jibe, etc.:** Sometimes, personal expressions can be very direct and unambiguous. However, more often than not they are ambivalent and as such very problematic

for NLP tasks. Similar to slang and abbreviations there are also language and demographically dependent and therefore very difficult to process by computational means;

- **Data Volumes:** While individual messages are relatively small, the number of messages published at any given moment in time can be very large. Analyzing them in real-time can pose significant challenges to the analyzing platform. For instance, according to Twitter, in 2011 approximately 1 billion tweets were sent every week (http://blog.twitter.com/2011/03/numbers.html);
- **Accuracy vs. Performance:** In general, an increased accuracy requires a more detailed analysis and therefore decreases the overall performance. Finding the right balance between the required accuracy and the best performance is a key challenge in machine learning and beyond that requires more adaptive and ideally self-adapting algorithmic principles.

RELATED WORK

Sentiment analysis and opinion mining has been around for a number of years, e.g. (Liu et al., 2003) have introduced textual affect sensing to analyze various information sources such as reviews, news articles, conversations, blogs, etc. However, in recent years it has become a very active area of research, which is mainly due to the rising popularity of real-time micro blogging platforms such as Twitter, which provide a very large and real-time pool of information that is very rich in personal opinion. In fact, Bifet et al. (2010) have referred to Twitter as a tool that reflects "what's-happening-right-now". Moreover, Jansen et al. (2009) discussed the impact of micro blogging as a form of electronic word-of-mouth branding that can provide new opportunities for companies. They reported that a significant number of blogs contained references to brand products also expressing positive or negative opinions about the product or the associated brand. This clearly shows that analyzing such sources can provide an important insight into the public opinion in almost real time, which reflects a valuable feedback mechanism that can be utilized for various purposes.

Probably the most popular and simplistic approaches to sentiment analysis is based on keyword spotting where specific sentiment bearing lexicons are used to determine the sentiment of a given input. Such lexicons contain keywords that correspond to the different sentiment categories (Liu et al., 2003). Most publicized algorithms only differentiate between positive and negative sentiment categories also assuming a neutral category if no sentiment is detected. Other approaches include more complex NLP or machine-learning algorithm that take into account the semantic structure of a phrase of a sentence or even larger constructs at e.g. document level.

One of the key challenges for this type of algorithm is the generation of generic, domain or topic specific or even multi-lingual lexicons that are applicable for any content. The most basic approach is to manually populate such indicators or to use online resources such as Wordnet (http://wordnet.princeton.edu/). Other approaches extract relevant indicators from pre-tagged information in an attempt to automatically generate relevant keyword sets. Anyhow, the impact of different sets of keywords have on the classification process is significant and with no clear evaluation available, further research is required to quantify the impact with respect to their domain as well as topic specificity.

A comprehensive survey of early research in sentiment analysis has been provided by Pang et al. (2008) which also discusses various techniques for an opinion oriented information retrieval system. Another good survey for the detection of sentiments in reviews has been presented by (Tang et al., 2009). Yang et al. (2007) researched the use of support vector machine (SVM) and conditional random fields (CRF) for the emotion-based classification of web blogs. In their work they trained each classifier on a sentence level and then applied it at document level. They showed that the CRF classifier outperformed the SVM but also recognized that the

last sentence of each document had the highest impact on the accuracy of the classification process. Similarly, Pang et al. (2002) used an SVM classifier on unigram features to classify movie reviews achieving an accuracy of over 80%. Read et al. (2005) used a set of emoticons to construct a domain and topic independent training data set for sentiment classification. He trained Naïve Bayes classifier and an SVM based classifier and achieved a mean accuracy of 61.5% and 70.1%, respectively.

Similar to this work, Pak et al. (2010) collected a corpus of approx. 300,000 tweets from Twitter, which was evenly split containing positive, negative or no emoticons. Based on this, three different classifiers SVM, CRF and Naïve Bayes were built with the latter yielding the best results. In this work a number of strategies were investigated to improve accuracy. Firstly, an entropy measure was introduced that relates to the probability distribution of the appearance of a given n-gram in different data sets. In this context a low entropy value points to a higher impact on the classification procedure. Thus discarding n-grams that have an entropy value above a given threshold will increase overall accuracy. Secondly, a "salience" value is computed for each n-gram, which indicated the relative prominence an item has in comparison to its neighboring items. In this case, a low value indicates lesser importance such that n-grams below a given threshold may be discarded.

Jiang et al. (2011) proposed a target-dependent Twitter classification mechanism that incorporates target or to be more accurate topic specific features as well as topic related tweets into the classification procedure. The former incorporates the features of words that are syntactically related whereas the latter is a context aware related approach that includes the sentiment labels of related but already classified tweets.

Wilson et al. (2005) have investigated the contextual polarity at phrase level. The work utilized a lexicon of positive and negative clues (or keywords) to classify the each phrase. However, unlike other approaches it first determines if a phrase is neutral or polar and only of it is polar an in depth analysis is performed to determine its polarity. As such it first classifies a given input to be either neutral or polar and if the latter it further classifies it to be either positive, negative or both. The latter in particular is of interest as most other approaches regard a given content as neutral if it includes both positive as well as negative sentiments. This work was later extended (Wilson et al., 2009) to include additional features providing a more detailed analysis. In particular the distinction between prior and contextual polarity has been investigated.

As pointed out in Banea et al. (2001) there is a growing need for multilingual subjectivity and sentiment analysis which is based on the fact the about two third of online content is not based on the English language. While most keyword-based classifiers simply require different lexicons to adapt to different languages, a more in depth analysis would require significant changes to adapt to language specific semantics. In this work relevant research was discussed for the development of language specific resources and methodologies to support multilingual sentiment analysis that focused on word and phrase level annotations, sentence labeling as well as document level annotations. Similarly, Pak et al. (2010) proposed a mechanism that disambiguates ambiguous adjectives in Chinese and classifies them into positive and negative sentiment polarity. The resulting classifier is language independent and does not require any additional user input.

KEYWORD-BASED SENTIMENT MINING

This section discusses a prototype algorithm that extracts various sentiments from short message based content such as reflected by micro blogs. Although it is possible to analyze larger input such as paragraphs or whole documents, the algorithms has not yet been evaluated in this context.

The current prototype as discussed next only differentiates between two categories, positive and negative or neutral in case no sentiment has been expressed. Either sentiment category is defined by a corresponding set of keywords as well as emoticons. Such that each sentiment category $S = (s_p, s_n, s_{p\_Em}, s_{n\_Em}, s_{not})$, where each $s = (K_1, K_2, \ldots, K_m)$, were K = ($k$, w) with k being the keyword or emoticon and w is the associated weight which is positive for all instances.

Prototype

As depicted in Figure 1, the developed prototype is composed of a number of individual modules capable of computing specific tasks. Each module incorporates a flexible publish / subscribe mechanism to which other modules can (un-)subscribe in order to retrieve the data that are published by a particular component. This mechanism allows for the flexible construction of the algorithm at runtime and also promotes decentralization as required for cloud based or multi-core environments.

The objective of the Pre-Processor module is to improve the quality of the input by removing noise and by pre-parsing it into distinct tokens that are ready for further processing. For example, repeating characters or words may be removed or known abbreviations are expanded. This task also allows combining individual tokens into bi- or multi-grams in the case a sentiment is expressed by multiple words. The thread pool module synchronizes the concurrent use of resources as well as the processing and distribution of individual content, which is buffered on input and subsequently assigned to individual processor modules on a request basis. Within the thread pool each pre-processor - sentiment processor module is executed as an independent thread. However, the load balancing of the threads itself depends on the underlying JVM as well as on the operating system. The Reader and Outputter modules are self-explanatory an will not be discussed whereas the remaining three modules can be summarized below.

The Sentiment Processor determines the sentiment of a given content as specified in Figure 2. For that it tests each token if it is contained in the set of keywords specified. If found the positive or negative weight is assigned to the respective token and the sentiment is determined as follows:

$$s = \frac{\sum w_p}{c_n} - \frac{\sum w_n}{c_p}$$

with w_p being the sum weight of all positive token found and w_n being the sum weight of all negative tokens, c_n being the number of negative

Figure 1. Module overview and process flow

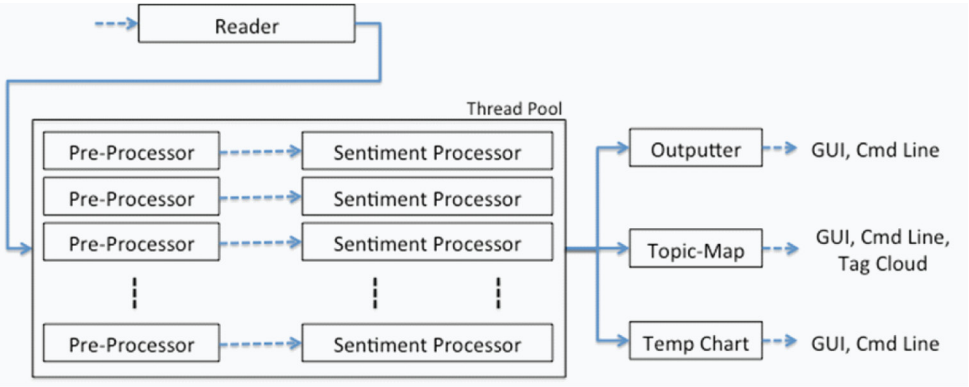


Figure 2. Sentiment processor

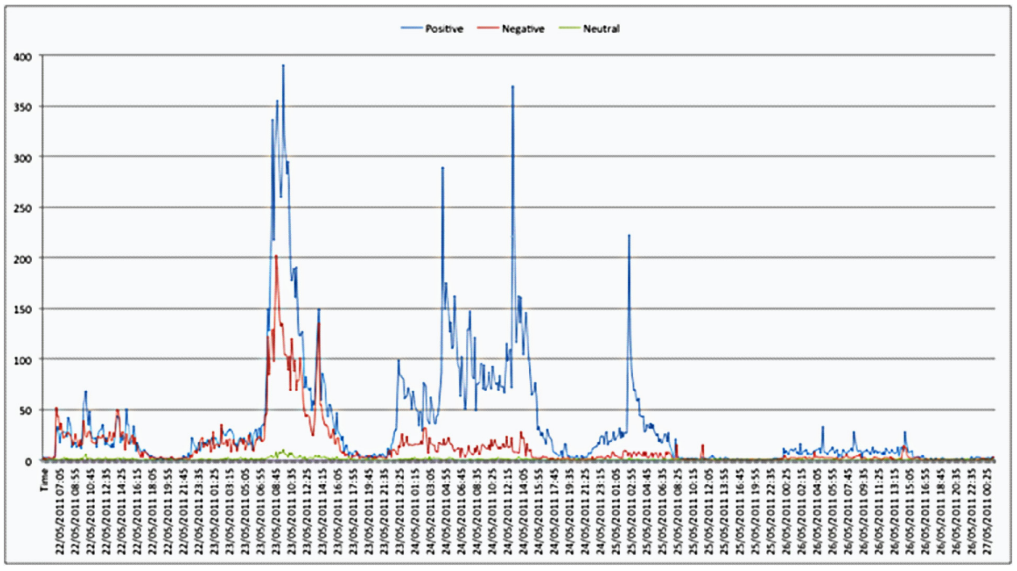
DetermineSentiment

set up pre-processing rules
set up sentiment categories S
set up topics base P // list of topic's
set up time windows K // list of windows of size k_s
→ Tweet t
    w_p = SUM of all positive weights $w \in s_p$
    w_n = SUM of all negative weights $w \in s_n$
    c_p = number of all positive keywords $k \in s_p$
    c_n = number of all negative keywords $k \in s_n$
  // Note: If a keyword is preceded by a NOT Keyword then the polarity of the keyword is reversed.
    IF $(w_p/c_n - w_n/c_p) = 0$
        w_p = SUM of all positive emoticons $w \in s_{p\_Em}$
        w_n = SUM of all negative emoticons $w \in s_{n\_Em}$
        c_p = number of all positive emoticons $w \in s_{p\_Em}$
        c_n = number of all negative emoticons $w \in s_{n\_Em}$
    end IF
$s = (w_p/c_n - w_n/c_p)$

keywords found and c_p being the number of positive keywords. If no keywords are found or if $s = 0$ then the algorithm searches for emoticons in the same way. Thus, emoticons will be ignored if the keyword based analysis yields a result. The reasons for this will be discussed in the next section on Performance and Evaluation. Based on the above calculation, each tweet is then classified to be either neutral for $s = 0$, negative for $s < 0$ and positive for $s > 0$. After the sentiment has been computed, the tweet is published to all registered components for further processing.

The Temp Chart module allows a more detailed view of how the sentiment has changed over time. For that sentiments are grouped into specific time periods as specified by the user. Figure 3 visualizes such a result clearly highlighting specific peaks of activity and changing sentiment.

Figure 3. Sentiment timeline



The Topic Map module reflects another mechanism to explore the individual features the overall sentiment is based upon. It shows a tree based sentiment visualization that is based on individual keywords or keyword clusters.

This provides a detailed overview of how individual topics or features influence the overall sentiment and as such enables the interpretation of the extracted sentiment in relation to its individual features. For instance, the results shown in Figure 4 relate to a well-known English footballer with strong sentiments expressed for the terms "injunction" and "footballer". However, while the overall sentiment and the sentiment for the term "injunction" was largely negative, the sentiment for the term "footballer" was positive. This clearly shows that sentiments are context sensitive, which needs to be interpreted on a feature level rather than on a general topic basis.

Multiple Sentiment Categories

The above only allows distinguishing between two categories. As discussed in the second section, this approach is too simplistic to capture the various sentiments humans can express. This section outlines an extension to the discussed algorithm, which is currently being implemented. In particular, it allows specifying multiple sentiment categories that are each defined by a corresponding set of keywords as well as emoticons. Such that $S = (s_1, s_2, \ldots, s_n)$, where $s = (K_1, K_2, \ldots, K_m)$, were $K = (k, w)$ with k being the keyword and w is the associated weight. Nevertheless, instead of explicitly classifying each message into a distinct category it provides a weight factor for each category indicating the strength of the corresponding sentiment. As such, a single content object may impact multiple sentiment categories. In this case the sentiment processor module computes the sentiment vector of a given content object. For that it tests for the occurrence of each token in the sets of keywords specified. If found the weight of the keyword is assigned to the respective token and the strength S for each specified sentiment category is calculated by $S_i = \left(\sum w\right) / n$, where i reflects the sentiment category, w is the weight assigned to the token and n the number of tokens found for that

Figure 4. Topic map and corresponding tag cloud (created with http://tagcrowd.com/)

specific category. The resulting sentiment vector reflects the various degrees of sentiment expressed in the content object. The category with the highest strength obviously reflects the most dominant sentiment, which allows for a distinct classification required. If multiple categories have the same strength than the number of tokens could be used additionally to determine the prevailing sentiment. In this case the category where most tokens have been found should take precedence. The overall sentiment towards a specific topic can be determined in the same way by averaging the sentiment vectors of the whole input set.

One of the problems in this context is the generation of the different keyword lexicons for each sentiment category. Some keywords may overlap or depend on different contexts or topics. As such the context specific adaptation of such keyword lexica needs to be further researched.

PERFORMANCE AND SCALABILITY

Overall performance and scalability are important aspects of any algorithm but they are particular important when dealing with potentially very large data volumes such as produced by todays social networking platforms. This section analysis the performance of the developed prototype and assess its scalability with respect to an increased data input and an increased set of keywords. Additionally, the multithreading aspect of the prototype will be discussed.

In order to assess the performance behavior of the algorithm, a data set containing 100,000 tweets has been extracted from Twitter using the provided API. Based on this, 10 data sets have been compiled which reflect the original data set times *n*. This replicative approach ensures that the runtime evaluation is only influenced by the number of records and not by the variation in content. Hence, the input size for each data set ranges from 100,000 tweets to 1,000,000 tweets. All experiments have been performed on a 2.53GHz Intel Core 2 Duo laptop with 4 GB of RAM running MacOS version 10.6.8. Also, to allow for comparison only two sentiment categories are used, positive and negative with an additional neutral category if no sentiment is detected.

The performance of the algorithm with respect to an increased data set is shown in Figure 5. As can be seen, the runtime ranges from less than 5 seconds to just over 40 seconds increasing in a linear fashion with respect to the size of the input data. This reflects an average of approximately 25,000 tweets (see Figure 6) per second.

Figure 6 shows the runtime behavior for an increased number of threads (1, 10, 100 threads). As shown, the runtime differs only slightly with

Figure 5. Runtime, number of tweets

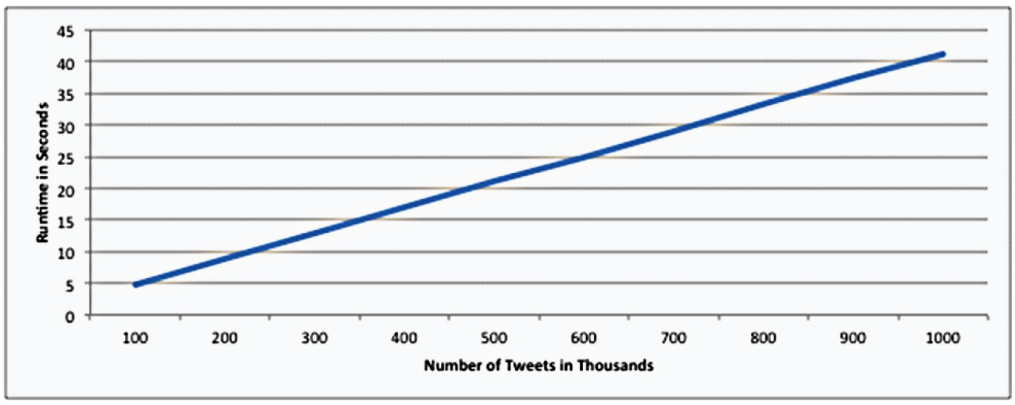
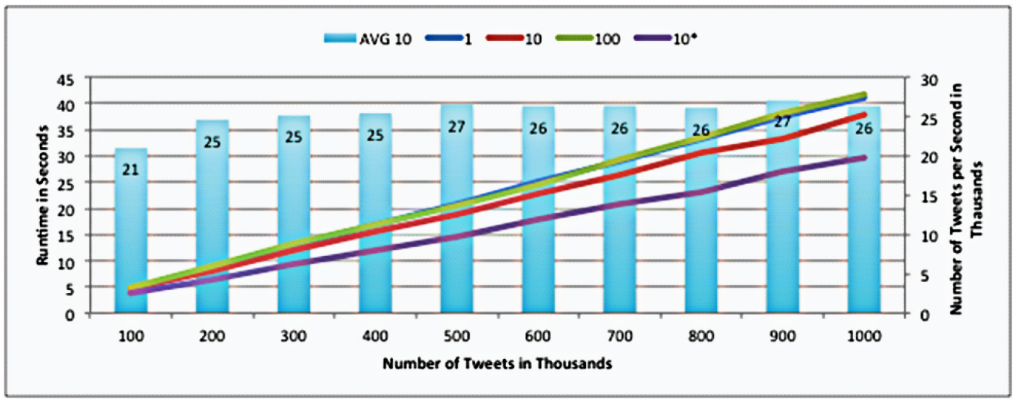


Figure 6. Runtime, number of threads

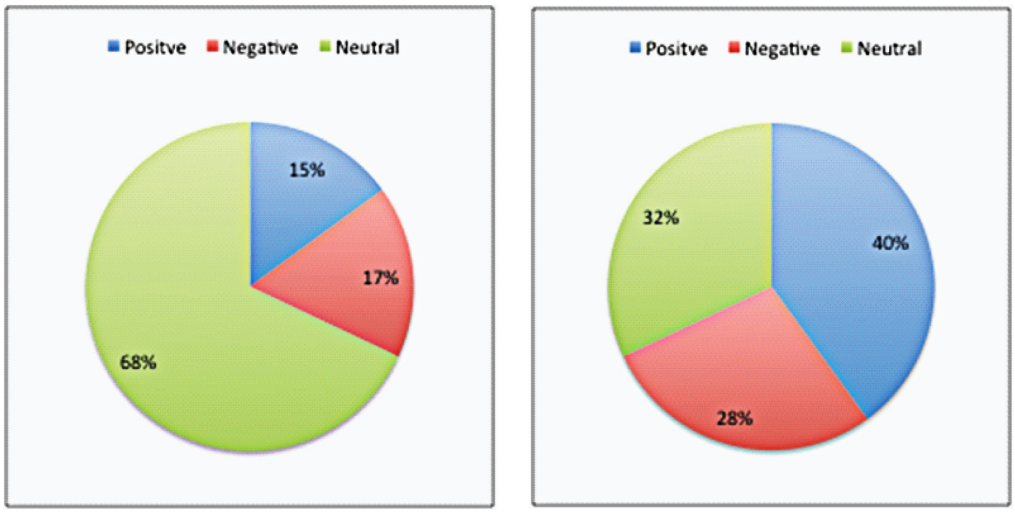


the difference becoming more significant with an increased number of tweets. The experiments showed that the runtime of the algorithm was improved if multiple threads were used. However, if more than 10 threads were used then the performance decreased again. This behavior was expected for two reasons. Firstly, when using multiple threads an additional overhead is required to synchronize the computation for each tweet. This overhead becomes more significant when the number of threads used is increased. Secondly, the test machine only has 2 cores such that a larger number of threads (more than 2) is actually irrelevant. The main reason that the performance increases if up to 10 threads are used is that the data I/O mechanisms are separated from the actual sentiment computation and as such finishes earlier compared to a single thread mechanisms. Nevertheless, it is expected that using multiple threads will reduce runtime significantly if the algorithm is executed on multi-core platforms that comprise larger number of processors. This however, needs to be further evaluated.

Another aspect to be evaluated is the size of the keyword set used as this dictates the number of steps required to determine the sentiment of a given tweet. In general, a larger set of keywords will lead to an increased runtime. Figure 6 shows the execution times for two different sets of keywords, 10 - approximately 6500 keywords including ca. 80 emoticons and 10* - 220 keywords including 14 emoticons. As expected, the runtime for the smaller set of keywords is better also becoming more significant when the size of the input set is increased. Nevertheless, considering the size of the different sets of keywords the impact they have on the overall runtime is well within acceptable limits. Moreover, when evaluating different results generated then it becomes clear that it is not the runtime that is important but the impact they have on the classification. Figure 7 shows the sentiment classification for the two sets of keywords for the test data set and as can be seen, the shorter set of keywords is less likely to distinguish between positive and neutral or between negative and neutral respectively. In general, a smaller set of keywords that uses only very distinct positive and negative indicators would produce less false positives or negatives, respectively, but would also be more likely to produce more false neutrals. Vice versa, a larger set of keywords is more likely to trigger a positive or negative classification, which would entail an increased risk of false positives or negatives, respectively.

Assessing the accuracy of the algorithm is difficult due to the simple fact that different persons may interpret the same tweet in different ways and as such may have different sentiments about them. Nevertheless, in order to quantify the accuracy, a publicly available data set[1] has been used which is also discussed in Pak et al.

Figure7. Sentiment distribution



(2010). The data have been classified manually into negative, neutral and positive. The results of this exercise are shown in Table 1 along with the original classification provided by Pak et al., As can be seen the manual classification differ by 9% and 18% respectively when compared to the original data. Out of necessity, we assume the Original to be correct.

Table 2 shows the accuracy of the algorithm compared to the original data set using three, slightly different evaluation criteria. Firstly, the data set was evaluated by including / excluding emoticons into the evaluation process. Next, a mixed mode has been used where emoticons were only used if no keywords where found in the current tweet.

As can be seen, the results differ significantly (> 10%) depending on the use of emoticons. After manually analyzing this behavior it was found that sarcasms, mockery etc. was often expressed through negative keywords followed by positive emoticons. Whereas positive as well as negative messages where often expressed through neutral keywords followed by a respective emoticon. Although this behavior cannot be guaranteed for all data sets it is still likely that this configuration yields the best accuracy as it focuses on the actual content of a given message and not on the attached emoticons.

The comparable low accuracy of 77% is based on the fact that the algorithm cannot explicitly detect sarcasm and has also difficulties

Table 1. Manual classification

	Original	User1	User 2
Positive	108	108	96
Neutral	33	38	56
Negative	75	70	64
Correct	216	196	177
Accuracy	100%	91%	82%

Table 2. Accuracy

	Without Emoticons	With Emoticons	Mixed
Positive	112	116	114
Neutral	38	33	34
Negative	66	67	68
Correct	162	143	166
Accuracy	75%	66%	77%

in dealing with multiple topics. For instance, "I like Mike but don't like Thomas" would be categorized into neutral rather than positive for Mike and negative for Thomas. In order to address these limitations a more detailed language analysis would be required which is beyond the scope of this work. Also, such analysis would significantly increase the runtime of the algorithm as well as its computational requirements.

APPLICATIONS

The web in general but in particular the various micro blogging sites are becoming ever more valuable as a source for real time opinion mining. This is mainly due to the fact that the contributing community is constantly growing and that individuals produce and distribute content on the spot when and where it happens. This latter aspect in particular allows for the almost real-time analysis of events. However, detecting them accurately, analyzing their importance and, eventually, modeling the underlying sentiment over time is a challenging and inherently complex task. Potential applications include but are not limited to the following examples:

- **Product Management:** The release of new or updated products is often accompanied by an increased activity of users engaging in various social networking platforms discussing the product as a whole but in particular its features and drawbacks. Such activity often surges shortly before and after a product has been released providing an almost instantaneous feedback channel that reflects the level success or failure of the new release. Modeling the features of a given product such as hardware and software specifics also correlating the multifaceted sentiment of users to these features would reveal a very detailed overview about the likes and dislikes and could also include demographic or user-group specific peculiars which would otherwise require time and cost extensive surveys especially across multiple countries. Similar, the public profiles of celebrities could be analyzed in great detail in the same way also correlating their activity to other events or activities;
- **Policy Making:** Policy making on a national or even international level is a complicated issue where public opinion plays an ever more important role. However, including public opinion has been proven to be very difficult as the time span between proposals and implementation is often too short. Moreover, required surveys would be too complicated and too cost intensive to conduct. Extracting the features of individual policies and correlating the public sentiment would provide a quick-response overview about the pro and cons and could also reveal additional aspects that are of concern to the general public which would have otherwise not be included.

CONCLUSION AND FUTURE WORK

This paper reviewed the foundations and requirements of short message based sentiment mining and discussed a keyword based classifier that has the potential to extract a balanced, multi-sentiment based profile allowing for the fact that in the micro blogging arena, there will be variation in sentiments and in the degree such sentiment is expressed. While the current algorithm only distinguishes between two categories, the theoretical foundations have already been drawn-up to include additional sentiment dimensions. A performance study has been presented highlighting the scalability of the proposed algorithm in relation to various influencing factors. In addition, the accuracy of the algorithm has been evaluated which highlighted that the results differ significantly for different keyword sets which was, however, expected.

Future work will include a number of aspects. Firstly, the existing algorithm is currently being extended to include additional sentiment dimensions. This will also require the development of context specific lexicons to allow for the fact that for different topics or users certain terms may have different contextual influence on the sentiment. Moreover, such lexicons need to be further evaluated to assess the impact of social and demographic differences. Secondly, the inclusion of user specific aspects needs to be further evaluated to determine the nature of a user's opinion based on an analysis of their previous postings or their social network within the context of social networking mechanism. Finally, a semantic interpretation of the overall sentiment profile, its individual features and an explanation for changes in the profile that are based on time series analysis mechanisms to be able to identify and to track changes in sentiment.

REFERENCES

Banea, C., Mihalcea, R., & Wiebe, J. (2011). Multilingual sentiment and subjectivity. In I. Zitouni, & D. Bikel (Eds.), *Multilingual natural language processing*. Prentice Hall.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science,* Canberra, Australia (pp. 1-15). Springer.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*, 2169–2188. doi:10.1002/asi.21149.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the International Conference on Intelligent User Interfaces.*

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation* (pp. 19–21). 2-9517408-6-7.

Pak, A., & Paroubek, P. (2010). Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)* (pp. 436-439). Stroudsburg, PA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135. doi:10.1561/1500000011.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the Student Research Workshop at the 2005 Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, MI.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760–10773. ISSN 0957-4174, 10.1016/j.eswa.2009.02.063.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Joint Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing.*

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3). doi:10.1162/coli.08-012-R1-06-90.

Yang, C., Hsin-Yih Lin, K., & Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Proceedings of the IEEE / WIC / ACM International Conference on Web Intelligence (WI'07)* (pp. 275-278).

ENDNOTES

[1] http://www.stanford.edu/~alecmgo/cs224n/twitterdata.2009.05.25.c.zip

Matthias Baumgarten has worked for over ten years as a research fellow in the Faculty of Informatics at the University of Ulster. His specific expertise lies in the area of knowledge discovery, context aware computing, distributed knowledge management, networked intelligence, autonomous systems and wireless sensor networks. He has been involved in several national and European research projects on a technical as well as on an administration level. Having received his PhD in 2005, he has co-authored more than 50 publications and has also been involved in the preparation and submission of various national and international project proposals. Recently he has moved into industry as a Senior Data Scientist with RepKnight.

Maurice Mulvenna is Professor of Computer Science in the School of Computing and Mathematics at the University of Ulster. He is a chartered fellow of the British Computer Society and has published over two hundred peer-reviewed papers in his core research area of computer science. His main research interests include artificial intelligence, social & pervasive computing, and inclusive computing.

Niall Rooney has a BSc degree in Information Technology from Queen's University, Belfast (1992) and a MSc in Computation from Oxford University (1996), and a PhD in Informatics from University of Ulster (2008). Niall worked a number of years in the software industry prior to becoming a researcher at the University of Ulster. He is a current member of the Artificial Intelligence & Applications Research Group at the University of Ulster. His research interests include natural language processing, information retrieval, data mining, machine learning and case based reasoning and he has published 40 papers in these areas. His primary area of research is in machine learning approaches for textual analytics.

John Reid is an expert project manager with broad experience across numerous programmes, projects, methodologies and companies across the UK and Ireland. John's key strength is combining his technical background with his project management experience and stakeholder management to ensure that the projects he is responsible for managing are delivered on time, to budget and to customer expectations.